CLAIMS

We Claim:

1. A computational method for predicting intramolecular and intermolecular biopolymer interactions comprising a screening process which comprises:

(*i*) a standardization process;

(*ii*) a threshold determining process; and

(*iii*) a chain elimination process, wherein said screening process results in a set of association predictions for intramolecular and/or intermolecular interactions.

2. The method of claim 1 wherein the standardization process categorizes expected frequency tables and converts nonuniformly scaled scores into uniformly scaled scores.

3. The method of claim 1 wherein the threshold determining process determines thresholds based upon distribution of uniformly scaled scores.

4. The method of claim 3 wherein the thresholds are determined based upon a measure of location and a measure of dispersion.

5. The method of claim 4 wherein the measure of location is a median distribution and the measure of dispersion is an interquartile range of the median distribution.

6. The method of claim 5 wherein the threshold is equal to a multiple of the interquartile range past a third quartile range of the interquartile range.

7. The method of claim 6 wherein the uniformly scaled score is a -log P score having a -log P threshold determined by:

$$\text{-log P - Threshold} = 3QV(p) + IQR(p)*\text{P-MULT},$$

where 3QV(p) denotes the third quartile range for the -log P score; IQR(p) denotes the interquartile range for the -log P score; and P-MULT represents a constant.

8. The method of claim 7 wherein P-MULT has a default value of 2.

9. The method of claim 6 wherein the uniformly scaled score is a V threshold determined by:

V-Threshold=3QV(v) + IQR(v)*V-MULT,

where 3QV(v) denotes the third quartile range for the V score; IQR(v) denotes the interquartile range for the V score; and V-MULT represents a constant.

10. The method of claim 9 wherein V-MULT has a default value of 1.

11. The method of claim 1 wherein the chain elimination process groups scores greater than thresholds determined in the threshold determining process based upon shared, common positions, and then determines a pinnacle association for each group of scores.

12. The method of claim 2, wherein the standardization process converts nonuniformly scaled scores into uniformly scaled P scores.

13. The method of claim 12, wherein the P scores are calculated using equations

$$DF = (I-1)(J-1) \quad \text{and}$$

$$P_{DF} = f(Q) = \frac{Q^{\frac{DF}{2}-1} e^{\frac{-Q}{2}}}{2^{\frac{DF}{2}} \Gamma\left(\frac{DF}{2}\right)} \quad ,$$

where I and J represent dimensions of an actual frequency table; P represents a probability density for degrees of freedom, DF, as a function of Q, where $Q = \chi^2$, and χ^2 is Pearson's χ^2 statistic.

14. The method of claim 12, wherein the P scores are calculated using Exact

Statistical Methods.

15. The method of claim 12, wherein the P scores are calculated using estimated-exact statistical methods.

16. The method of claim 15 wherein the estimated-exact statistical method is based on a monte carlo simulation of a distribution.

17. The method of claim 13, wherein the expected frequency table is Normal.

18. The method of claim 14, 15, or 16 wherein the expected frequency table is Sparse.

19. The method of claim 1, wherein the threshold determining process is performed on -log of P scores.

20. The method of claim 1, wherein the threshold determining process is performed on V scores.

21. The method of claim 11, wherein the pinnacle association for each group of scores is chosen first according to highest P, then lowest DF, and then highest V.

22. A computational method for predicting intramolecular and intermolecular interactions comprising:

(a) obtaining genomic/biopolymer sequence data;

(b) subjecting the genomic/biopolymer sequence data to an alignment process, wherein said alignment process produces a set of sequence alignment data;

(c) subjecting the set of sequence alignment data to a combinatorial matching process, wherein said combinatorial matching process produces data indicating possible associations;

(d) preparing a set of actual frequency tables using the data indicating possible associations and the set of sequence alignment data;

(e) subjecting the set of actual frequency tables to a scoring process to obtain a set of scores; and

(f) subjecting the set of scores to a screening process to produce a set of association information, wherein said association information predicts intramolecular and/or intermolecular biopolymer interactions.

23. The method of claim 22 wherein the scoring process further comprises:

(*i*) generating a set of expected frequency tables from the set of actual frequency tables; and

(*ii*) comparing the expected frequency tables and actual frequency tables to generate a score.

24. The method of claim 23 wherein cells of the set of expected frequency tables have values determined by:

$$u_{ij} \approx \hat{u}_{ij} = \frac{n_{i+} n_{+j}}{n},$$

where $\hat{u}_{ij}$ (which is an approximation of u_{ij}) is an expected frequency of cells in row i and column j of the expected frequency table; n = sum of all counts; n_{i+} = total of cells in row i; n_{+j} = total of cells in column j.

25. The method of claim 23 wherein the comparing step is performed using a Pearson's A^2 statistic.

26. The method of claim 25 wherein the Pearson's A^2 statistic is represented by:

$$\chi^2 = \sum_{i,j} \frac{(n_{ij} - u_{ij})^2}{u_{ij}},$$

where n_{ij} = a value of a cell in row i and column j and u_{ij} is an expected frequency of a cell in row i and column j of the expected frequency table.

27. The method of claim 22 wherein the scoring process comprises subjecting the frequency tables to more than one scoring method.

28. The method of claim 27 wherein the scoring process further comprises obtaining a V score using a Cramer's V statistic.

29. The method of claim 28 wherein the Cramer's V statistic is represented by:

$$V = \sqrt{\frac{\chi^2 / n}{\min(I - 1, J - 1)}},$$

where A^2 is Pearson's A^2 statistic; n = the sum of all cell counts; I = number of rows; J = number of columns; and *min(I-1, J-1)* is a function which returns a lower of two values being compared, wherein A^2 is given by:

$$\chi^2 = \sum_{i,j} \frac{(n_{ij} - u_{ij})^2}{u_{ij}},$$

where n_{ij} = the value of a the cell in row i and column j and u_{ij} is an expected frequency of a cell in row i and column j of the expected frequency table.

30. The method of claim 22 wherein the screening process comprises:

(*i*) a standardization process;

(*ii*) a threshold determining process; and

(*iii*) a chain elimination process, wherein said screening process results in a set of association predictions for intramolecular and/or intermolecular interactions.

31. The method of claim 30 wherein the standardization process categorizes expected frequency tables and converts nonuniformly scaled scores into uniformly scaled scores.

32. The method of claim 30 wherein the threshold determining process determines thresholds based upon distribution of uniformly scaled scores.

33. The method of claim 32 wherein the thresholds are determined based upon a measure of location and a measure of dispersion.

34. The method of claim 33 wherein the measure of location is a median distribution and the measure of dispersion is an interquartile range of the median distribution.

35. The method of claim 34 wherein the threshold is equal to a multiple of the interquartile range past a third quartile range of the interquartile range.

36. The method of claim 35 wherein the uniformly scaled score is a -log P score having a -log P threshold determined by:

$$\text{-log P – Threshold} = 3QV(p) + IQR(p)*\text{P-MULT},$$

where 3QV(p) denotes the third quartile range for the -log P score; IQR(p) denotes the interquartile range for the -log P score; and P-MULT represents a constant.

37. The method of claim 36 wherein P-MULT has a default value of 2.

38. The method of claim 35 wherein the uniformly scaled score is a V threshold determined by:

$$\text{V-Threshold} = 3QV(v) + IQR(v)*\text{V-MULT},$$

where 3QV(v) denotes the third quartile range for the V score; IQR(v) denotes the interquartile range for the V score; and V-MULT represents a constant.

39. The method of claim 38 wherein V-MULT has a default value of 1.

40. The method of claim 30 wherein the chain elimination process groups scores greater than thresholds determined in the threshold determining process based upon shared, common positions, and then determines a pinnacle for each group of scores.

41. The method of claim 31, wherein the standardization process converts nonuniformly scaled scores into uniformly scaled P scores.

42. The method of claim 41, wherein the P scores are calculated using equations

$$DF = (I-1)(J-1) \quad \text{and}$$

$$P_{DF} = f(Q) = \frac{Q^{\frac{DF}{2}-1} e^{\frac{-Q}{2}}}{2^{\frac{DF}{2}} \Gamma\left(\frac{DF}{2}\right)} \quad ,$$

where I and J represent dimensions of an actual frequency table; where P represents a probability density for degrees of freedom, DF, as a function of Q, where $Q = \chi^2$.

43. The method of claim 41, wherein the P scores are calculated using Exact Statistical Methods.

44. The method of claim 43, wherein the P scores are calculated using estimated-exact statistical methods.

45. The method of claim 44 wherein the estimated-exact statistical method is based on a monte carlo simulation of a distribution.

46. The method of claim 43, the expected frequency table is Normal.

47. The method of claim 43, 44 or 45 wherein the expected frequency table is Sparse.

48. The method of claim 30, wherein the threshold determining process is performed on -log of P scores.

49. The method of claim 30, wherein the threshold determining process is performed on V scores.

50. The method of claim 40, wherein the pinnacle association for each group of scores is chosen first according to highest P, then lowest DF, and then highest V.

51. A computational method for predicting intramolecular and intermolecular interactions comprising:

(a) obtaining genomic/biopolymer sequence data;

(b) subjecting the genomic/biopolymer sequence data to a first alignment process wherein said first alignment process produces a first set of sequence alignment data;

(c) subjecting the first set of sequence alignment data to a combinatorial matching process wherein said combinatorial matching process produces data indicating possible associations;

(d) preparing a set of actual frequency tables using the data indicating possible associations and the set of sequence alignment data;

(e) subjecting the set of actual frequency tables to a scoring process to obtain set of scores;

(f) subjecting the set of scores to a screening process to produce a first set of association information, wherein said association information predicts intramolecular and/or intermolecular biopolymer interactions;

(g) subjecting the actual frequency tables corresponding to the first set of association information, to a misalignment process wherein said misalignment process produces misaligned sequence information;

(h) using the misaligned sequence information to realign sequences to produce a second set of sequence alignment data superseding the first set of sequence alignment data ; and

(i) repeating steps (c) through (f) in combination substituting the second set of sequence alignment data for the first set of sequence alignment data to produce a second set of association information, wherein said second set of association information predicts intramolecular and intermolecular biopolymer interactions.

52. The method of claim 51 wherein steps (g) through (i) are repeated at least one additional time.

53. The method of claim 51, wherein the misalignment process further comprises:

(i) generating an AR table for each actual frequency table;

(ii) categorizing cells in each AR table in a MISALIGNED category or a GOOD category wherein a cell of an AR table having an absolute value of less than or equal to a threshold value ARTHRESH is categorized in the MISALIGNED category and wherein a cell of an AR table having an absolute value of greater than a threshold value of ARTHRESH is categorized in the GOOD category;

(iii) using the cells in the MISALIGNED category to identify corresponding misaligned sequences,

(iv) using the cells in the GOOD category to identify suggested alternative alignments for the misaligned sequences, and

(v) utilizing the suggested alternative alignments to produce the second sequence alignment.

54. The method of claim 53, wherein each element AR_{ij} in row i and column j of an AR table is generated using the equation:

$$AR_{ij} = \frac{n_{ij} - \hat{u}_{ij}}{\sqrt{\hat{u}_{ij}(1-p_{i+})(1-p_{+j})}} \quad ,$$

where $p_{i+} = n_{i+}/n$; $p_{+j} = n_{+j}/n$; n = sum of all cell counts; n_{i+} = total of cells in row i; n_{+j} = total of cells in column j; n_{ij} = value of the cell in row i and column j; and $u_{ij} \approx \hat{u}_{ij} = \frac{n_{i+}n_{+j}}{n}$.

55. The method of claim 53, wherein ARTHRESH has a default value of 1.0

56. The method of claim 53, wherein three cells having highest AR values in each AR table are categorized in the GOOD category and predict molecular interactions that exist in nature at positions corresponding to the cells.

57. The method of claim 53, wherein a sequence is identified as misaligned if it contributes at least a specified number, REPEAT, cells to the MISALIGNED category.

58. The method of claim 57, where a default value for REPEAT is 2.

59. A computational method for predicting intramolecular and intermolecular biopolymer interactions comprising a misalignment process which comprises:

(i) generating an AR table for actual frequency tables of a first sequence alignment;

(ii) categorizing cells in each AR table in a MISALIGNED category or a GOOD category wherein a cell of an AR table having an absolute value of less than or equal to a threshold value ARTHRESH is categorized in the MISALIGNED category and wherein a cell of an AR table having an absolute value of greater than a threshold value of ARTHRESH is categorized in the GOOD category;

(iii) using the cells in the MISALIGNED category to identify corresponding misaligned sequences,

(iv) using the cells in the GOOD category to identify suggested alternative alignments for the misaligned sequences, and

(v) utilizing the suggested alternative alignments to produce a second sequence alignment.

60. A sequence alignment method comprising a misalignment process which comprises:

(i) generating an AR table for actual frequency tables of a first sequence alignment;

(ii) categorizing the cells in each AR table in a MISALIGNED category or a GOOD category wherein a cell of an AR table having an absolute value of less than or equal to a threshold value ARTHRESH is categorized in the MISALIGNED category and wherein a cell of an AR table having an absolute value of greater than a threshold value of ARTHRESH is categorized in the GOOD category;

(iii) using the cells in the MISALIGNED category to identify corresponding misaligned sequences;

(iv) using the cells in the GOOD category to identify suggested alternative alignments for the misaligned sequences; and

(v) utilizing the suggested alternative alignments to produce a second sequence alignment.

61. A sequence alignment method for aligning biopolymer sequences comprising:

(a) obtaining genomic/biopolymer sequence data;

(b) subjecting the genomic/biopolymer data to a first alignment process wherein said first alignment process produces a first set of sequence alignment data;

(c) subjecting the first set of sequence alignment data to a combinatorial matching process wherein said combinatorial matching process produces a first list of combinations of positions in biopolymer sequence data;

(d) preparing an actual frequency table for each of the combinations of positions in the first list of combinations of positions;

(e) subjecting each actual frequency table to a scoring process to obtain a set of scores;

(f) subjecting the set of scores to a screening process to produce a second list of combinations of positions, wherein the second list of combinations of positions consists combinations of positions that are highly ordered.

(g) subjecting each actual frequency table prepared for each of the combinations of positions in the second list of combinations positions to a misalignment process wherein said misalignment process produces a list of misaligned sequence information, and wherein said list of misalignment information is an indicator of alignment quality; and

(h) using the misaligned sequence information to realign sequences to produce a refined sequence alignment.

62. The method of claim 61 wherein steps (c) through (g) are repeated in combination at least once.